On Decoder-Only Architecture For Speech-to-Text and Large Language Model Integration

Year: 2023 | *Citations: 181* | *Authors: Jian Wu, Yashesh Gaur, Zhuo Chen*

Abstract

Large language models (LLMs) have achieved remarkable success in the field of natural language processing, enabling better human-computer interaction using natural language. However, the seamless integration of speech signals into LLMs has not been explored well. The “decoder-only“ architecture has also not been well studied for speech processing tasks. In this research, we introduce Speech-LLaMA, a novel approach that effectively incorporates acoustic information into text-based large language models. Our method leverages Connectionist Temporal Classification and a simple audio encoder to map the compressed acoustic features to the continuous semantic space of the LLM. In addition, we further probe the decoder-only architecture for speech-to-text tasks by training a smaller scale randomly initialized speech-LLaMA model from speech-text paired data alone. We conduct experiments on multilingual speech-to-text translation tasks and demonstrate a significant improvement over strong baselines, highlighting the potential advantages of decoder-only models for speech-to-text conversion.